Aalto University
School of Science

Quality of Analytics as an Approach for Optimizing ML Systems: Initial results and roadmap

Hong-Linh Truong,
Department of Computer Science
https://rdsea.github.io

Acknowledgement

- **Include results from joint works with**
  - Matt Baughman, Nifesh Chakubaji, Kyle Chard, Ian Foster (University of Chicago)
  - Krists Kreics (master thesis with Sellforte)
  - Minjung Ryu (master thesis with Solibri)
- **Note: work in progress**

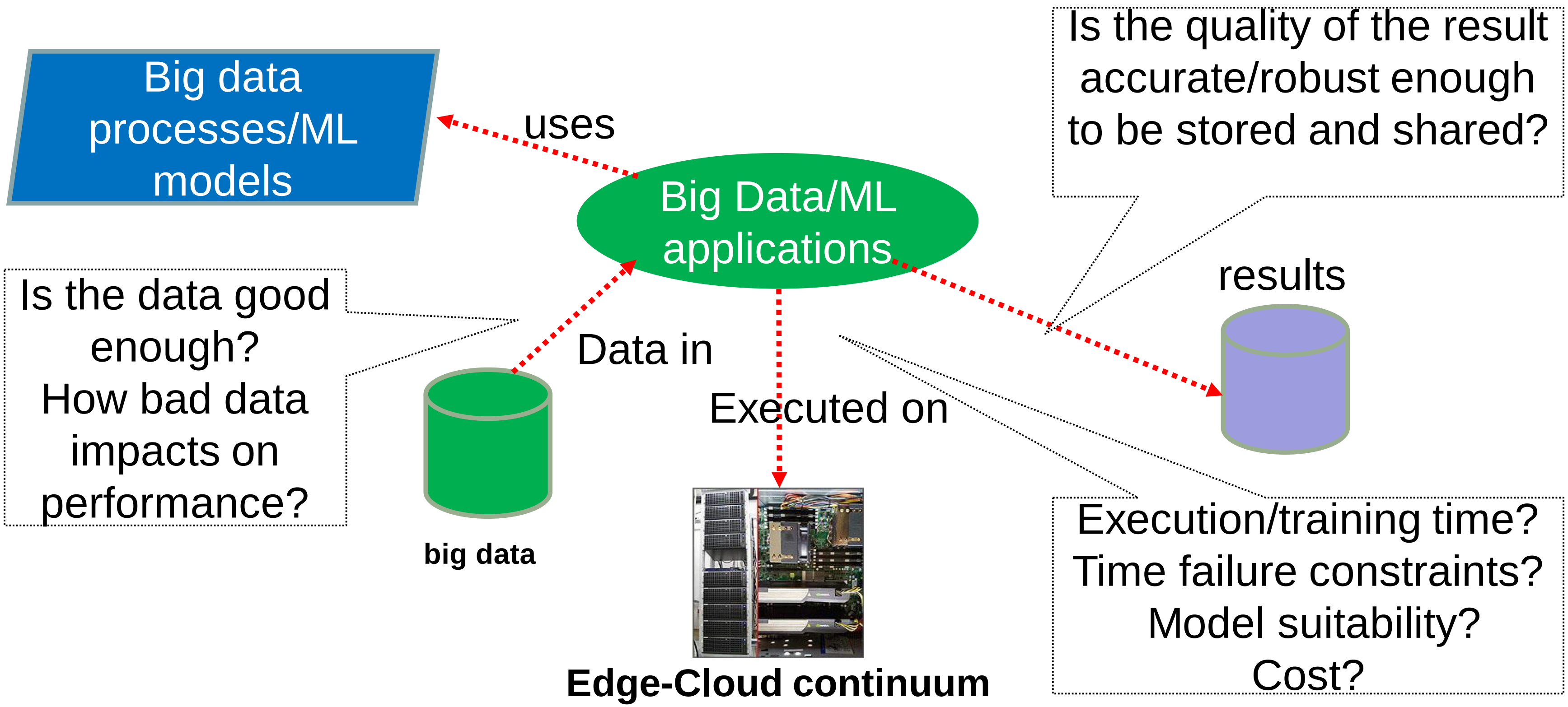
Content

- **Quality of Analytics (QoA) and Principles of Elasticity**
- **QoA-aware optimization for ML systems**
- **Initial results**
- **Next steps and conclusions**

Research focus: optimize the end-to-end ML pipelines

- **Building end-to-end ML (for production) is hard**
  - Several phases from data collection to training to model serving
- **The "system aspect" for ML**
  - Managing dymamic computing and data resources
  - Making ML models serving under "AI as a service" robust, reliable and resilient
- **Optimization from the software systems view**
  - Beyond ML Benchmarks and hyperparameter optimization
  - End-to-end runtime management, ML model serving, and ML experiments

Quality of Analytics (QoA)



QoA

- **Challenges in managing quality across multiple data analytics contexts (DACs).**
  - Interactions with data processing/ML frameworks
  - Interactions with different input and output data sources
  - Interactions with different system services for provisioning, monitoring and control
- **QoA as a composition of multi-dimensional data quality, performance, cost, etc.**
  - QoA as a contract: a "reliable service" should guarantee expected QoA from customers

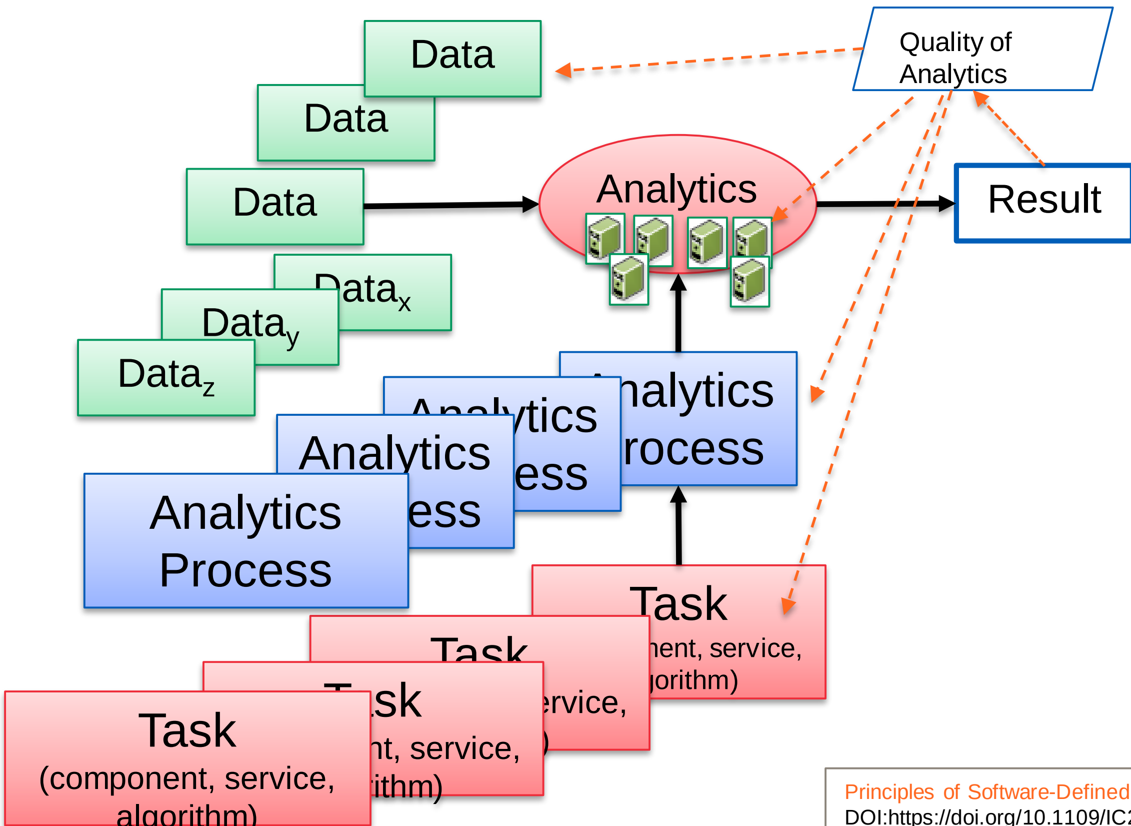
Hong-Linh Truong, Aitor Murguzur, and Erica Yang. 2018. Challenges in Enabling Quality of Analytics in the Cloud. J. Data and Information Quality 9, 2, Article 9 (January 2018), 4 pages. DOI:https://doi.org/10.1145/3138806

Principles of Elasticity

Ability to stretch the "form" under "pressure/force" and return to the normal shape

- **Demand elasticity**
  - Elastic demands from users/customers
- **Output elasticity**
  - Multiple outputs with different price and quality models
- **Input elasticity**
  - Elastic data inputs, e.g., deal with opportunistic data
- **Elastic quality models associated resources and processes**

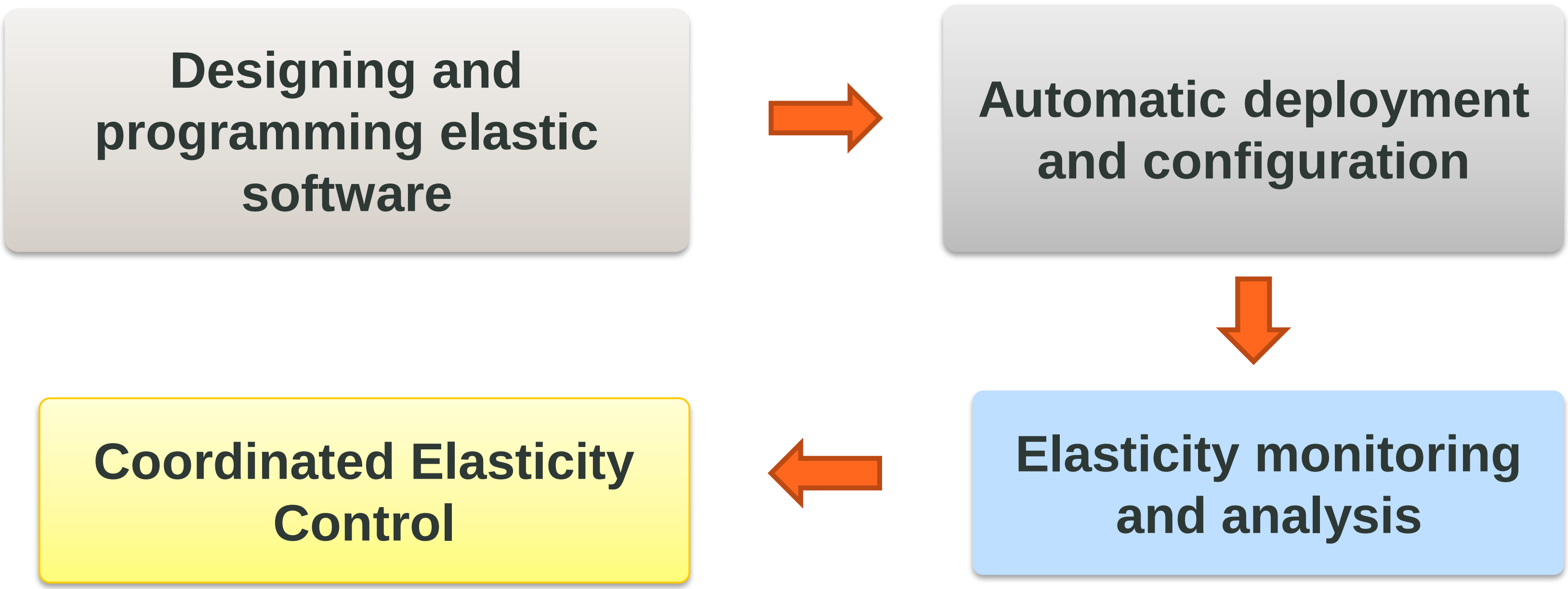
QoA and Multi-dimensional elasticity



- **More data → more compute resources (e.g. more VMs)**
- **More types of data → more, different tasks → more analytics processes**
- **Change quality of analytics**
  - Change quality of data
  - Change response time
  - Change cost
  - Change models and their quality

Principles of Software-Defined Elastic Systems for Big Data Analytics.
DOI:https://doi.org/10.1109/IC2E.2014.67

Elasticity engineering



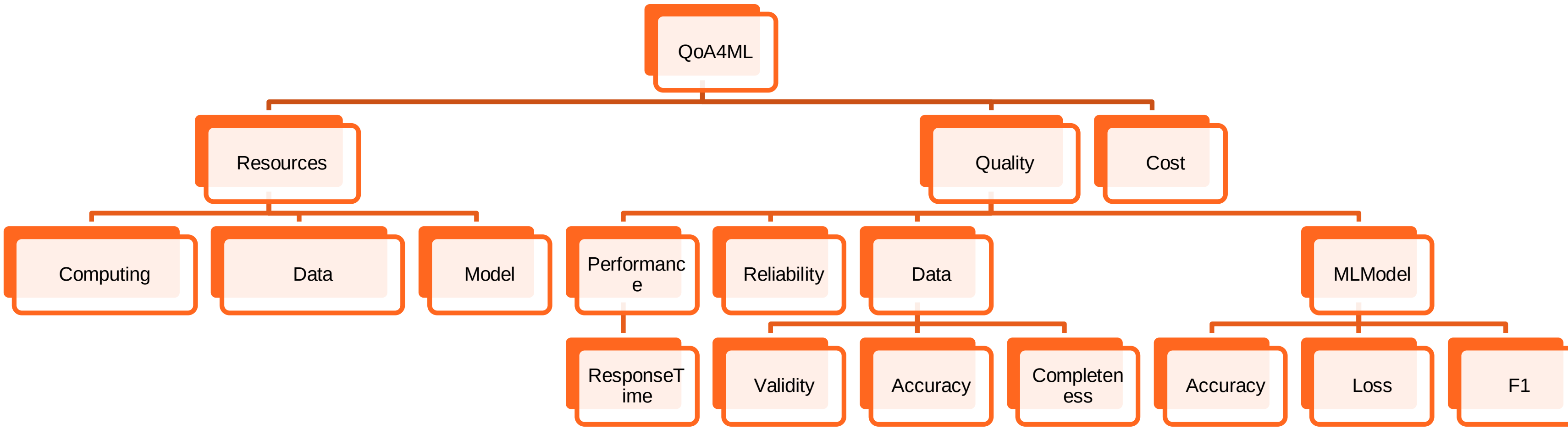
QoA approach for ML systems

QoA as a contract for optimizing ML

- **Quality of analytics: a complex relationships between quality of results, performance and cost**
  - Quality of results are characterized by the users/domain expert, e.g., quality of data of the output, accuracy of the model
  - Inputs have complex characteristics: input data (quality of data, volume) and machines (e.g., computation)
  - Complex types of cost (money) and performance
- **QoA as a contract**
  - The optimization of ML systems is based on the specified QoA
  - Runtime changes and updates by people or intelligent software

Determining most critical elasticity dimensions for ML systems

Example of dimensions

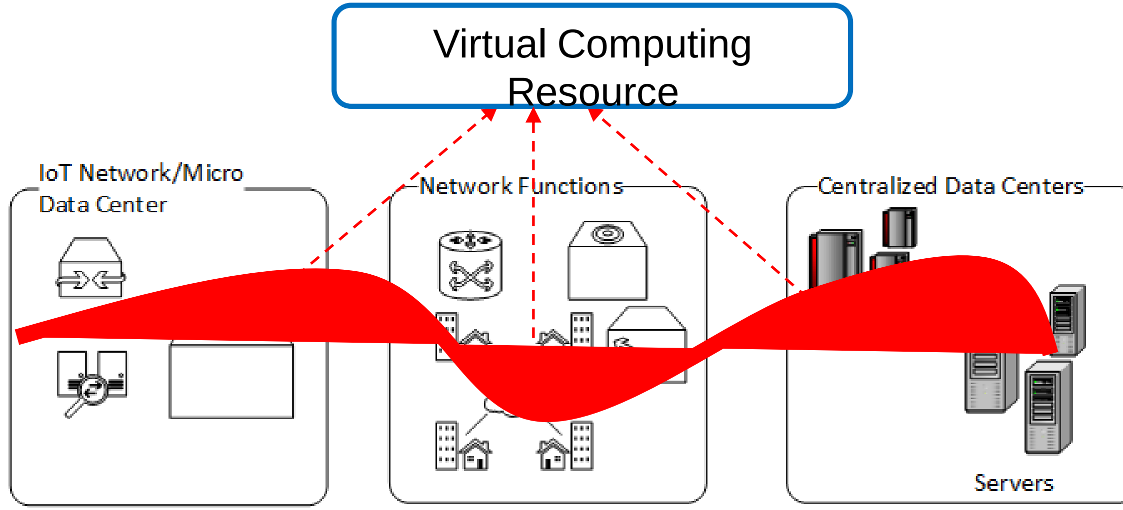


Elasticity engineering for ML

- **Conceptualizing and modeling elastic objects**
  - ML models, computing resources, data and QoA metrics
- **Defining and capturing elasticity primitive operations**
  - Change resources, QoA metrics, model parameters, input data
- **Programming features for elastic objects**
  - With ML flows, coordinating QoA adjustment, dynamic serving models
- **Runtime deploying, control, and monitoring techniques for elastic objects**

Elastic computing resources

Application-specific Resource Ensemble (ASRE)



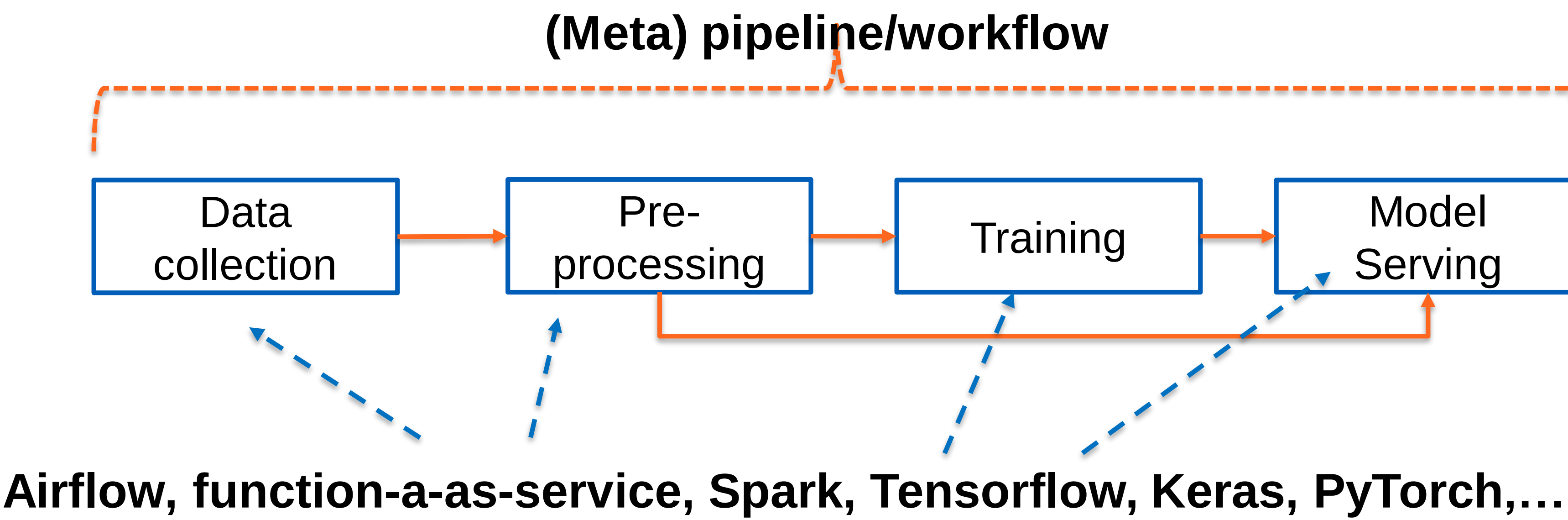
Coral with Edge TPU System-on-Module, Google Edge TPU ML accelerator coprocessor

Jetson NVIDIA (GPU+CPU)

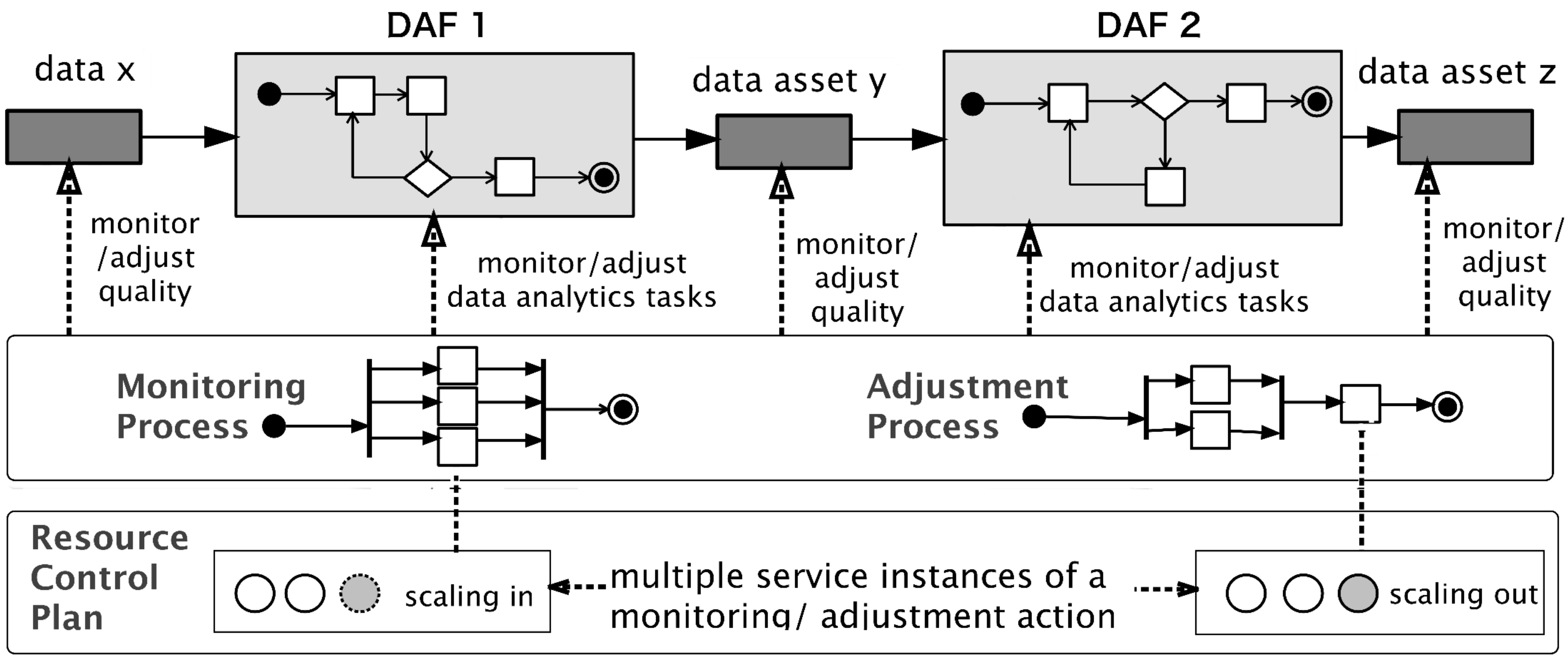
Hong-Linh Truong, ASRE – Application-specific Resource Ensembles across Edges and Clouds, Working paper, 2019

Elastic objects in ML workflows

- **Multiple levels:**
  - Meta-workflow or -pipeline
  - Inside each phase: pipeline/workflow or other types of programs



Elasticity primitive operations for Data Analysis Flows (DAF) model



Source: On Developing and Operating Data Elasticity Management Process. https://doi.org/10.1007/978-3-662-48616-0_7

Some initial results

With results from:

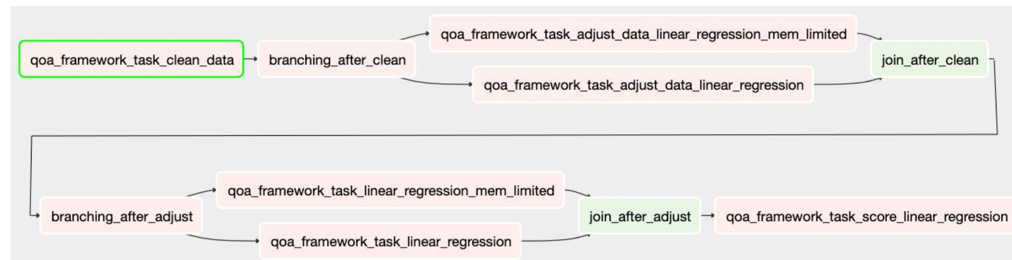
- Kreics Krists, „*Quality of analytics management of data pipelines for retail forecasting*,“, Aalto CS Master thesis, 2019, https://aaltodoc.aalto.fi/handle/123456789/39908
- Minjung Ryu, „*Machine Learning-based Classification System for Building Information Models* “, Aalto CS Master thesis, 2020 (to be finalized)
- Minjung Ryu, Linh Truong, „*Understanding Quality of Analytics Tradeoffs in an End-to-End Machine Learning-based Classification System for Building Information Modeling*”, 2020, Working paper.
- Matt Baughman, Nifesh Chakubaji, Hong-Linh Truong, Krists Kreics, Kyle Chard, Ian Foster, *Measuring, Quantifying, and Predicting the Cost-Accuracy Tradeoff,* IEEE International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications, IEEE BigData 2019, https://research.aalto.fi/files/38801332/paper.pdf

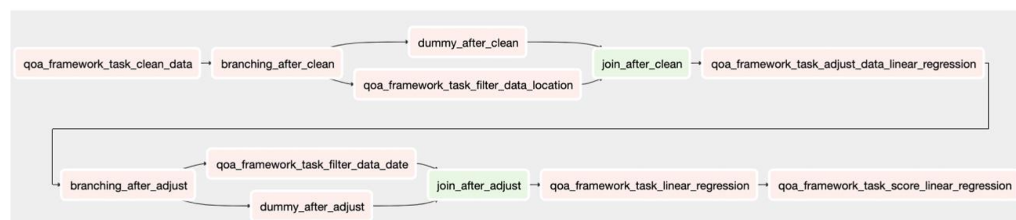
Industrial retail forecast (with Sellforte)

Forecast where to put marketing information, example of data

date	id	name	volume	price	cost	promo	category_net	margin	category1	category2	location	sales
07/01/2018	100	Chicken	38144.0	3.79	2.7	0	451692.0	0.25	Meat	Food	Helsinki	144565.76
14/01/2018	100	Chicken	36420.0	3.79	2.66	0	414342.0	0.25	Meat	Food	Helsinki	138031.8
21/01/2018	100	Chicken	35322.0	3.79	2.66	0	381854.0	0.25	Meat	Food	Helsinki	133870.38

- **Metrics:**
  - Data size, R square value, time, and cost
- **Pipelines**
  - Tune pipelines with QoA primitive actions

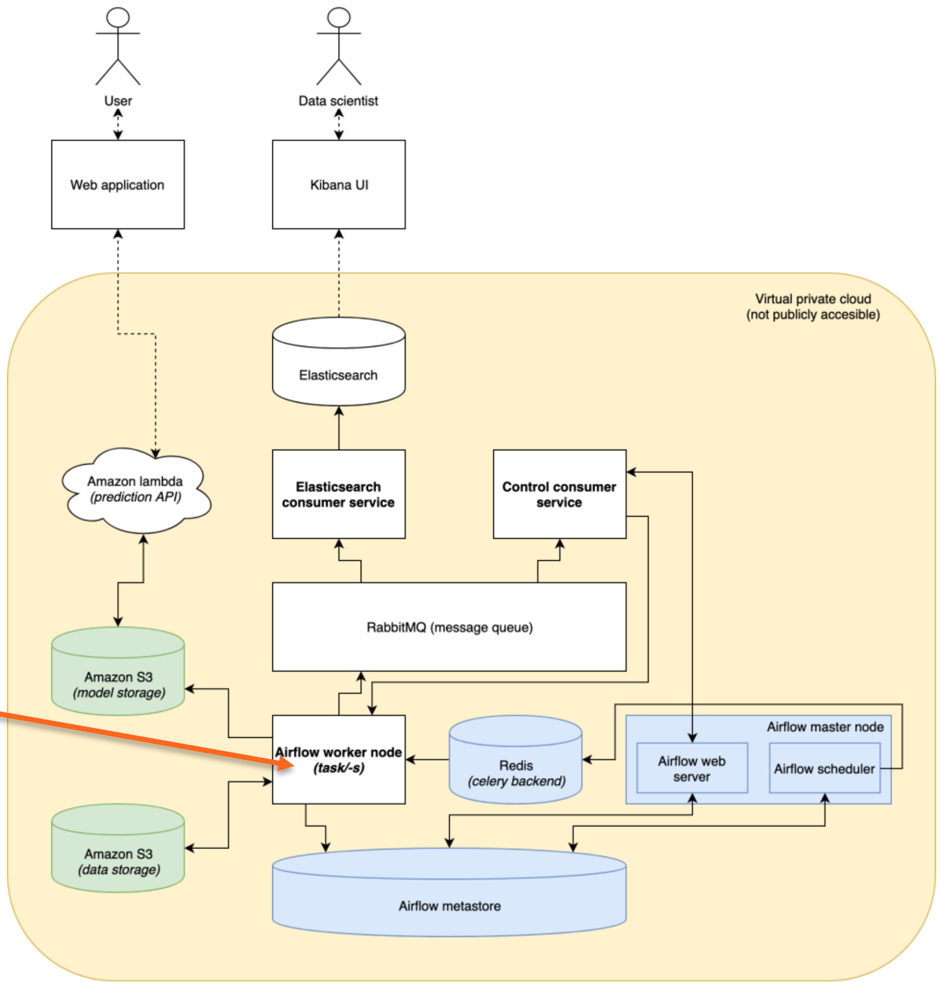




Source: Kreics Krists, „Quality of analytics management of data pipelines for retail forecasting,“, Aalto CS Master thesis, 2019

Industrial retail forecast (with Sellforte)

Monitoring various metrics, including user-defined quality of data



Source: Kreics Krists, „Quality of analytics management of data pipelines for retail forecasting“, Aalto CS Master thesis, 2019

Initial results

Custom cost function

```python
def get_fargate_metrics_object(cpu, ram, elapsed_time, previous_result):
    # Fargate service cost per second
    FARGATE_CPU_COST = 0.04048 / 60 / 60
    FARGATE_RAM_COST = 0.004445 / 60 / 60
    if previous_result and 'cost_usd' in previous_result:
        cpu_cost = previous_result['cost_cpu'] + FARGATE_CPU_COST
        ram_cost = previous_result['cost_ram'] + (ram['used']/1024/1024/1024) * FARGATE_RAM_COST
    else:
        cpu_cost = FARGATE_CPU_COST
        ram_cost = (ram['used']/1024/1024/1024) * FARGATE_RAM_COST

    return { 'cost_cpu': cpu_cost, 'cost_ram': ram_cost, 'cost_usd': ram_cost + cpu_cost }
```

Custom instrumentation for model quality

```python
# model_score returns a dict -> { 'r2_squared': r2_squared_score }
model_score = score_model(store, model, data_path, preset)
pm.log_analytics_metric(model_score)
```

Source: Kreics Krists, „Quality of analytics management of data pipelines for retail forecasting“, Aalto CS Master thesis, 2019

Examples of actions in Elasticity Primitive Operations

```python
def default_get_control_action(body_dict):
    index = body_dict.pop('metric_type', None)
    print(body_dict, flush=True)
    try:
        if index == 'metrics':
            if body_dict['cost_usd'] > 1 or body_dict['time_elapsed'] > 500:
                return 'SOFT_STOP'
            elif body_dict['time_elapsed'] > 1000:
                return 'HARD_STOP'

        elif index == 'data_logs':
            if body_dict['task_name'] == 'clean_data':
                if body_dict['in']['train.csv'] / 2 > body_dict['out']['train.csv']:
                    return 'SOFT_STOP'

        elif index == 'analytics':
            if body_dict['payload']['r2_squared'] < 0.2:
                return 'SOFT_STOP'
        else:
            print('No valid index found!')
            return -1
    except KeyError:
        pass
```

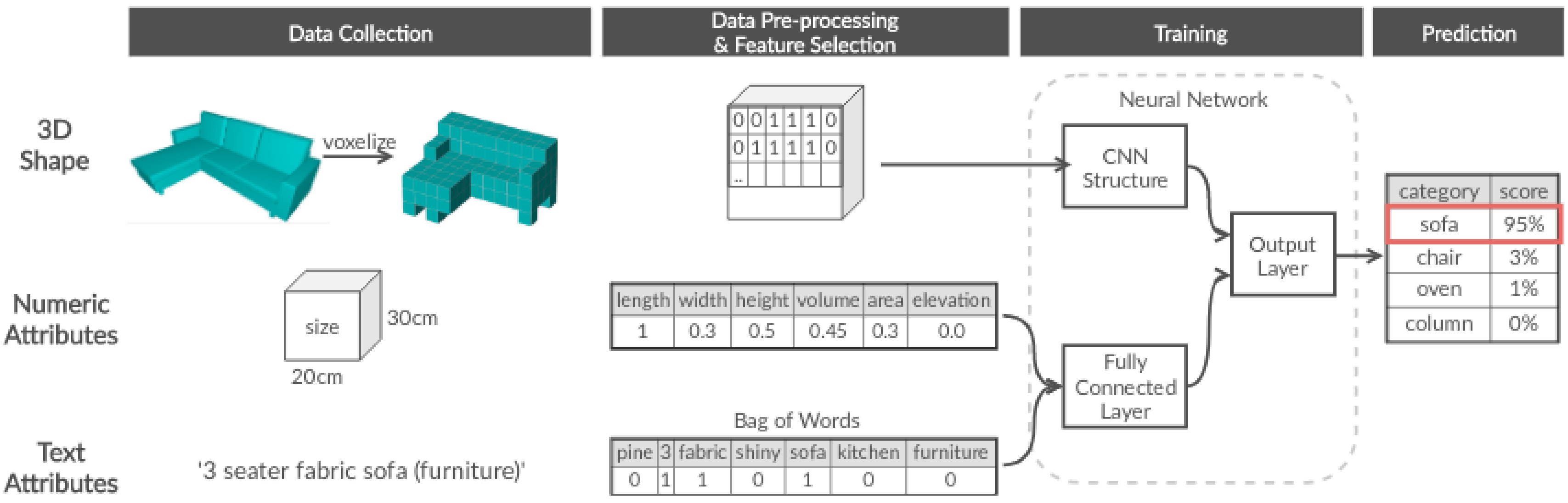
Initial results

- **Running with Airflows in Amazon EC2**
- **Apply different actions to change “store” (domain objects) and computing resources**
- **Real improvement (from the domain expert) with 1 million rows case**

  13.3% lower accuracy and 44% shorter time, R squared value was 9.5% lower → could good enough results for 50% of total store locations

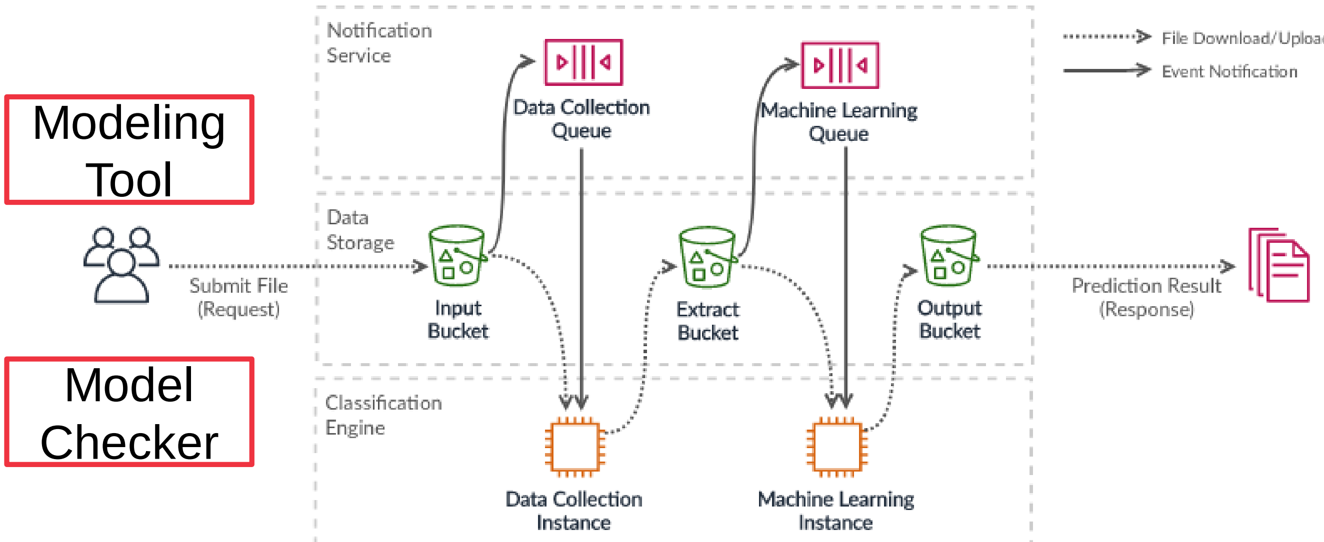
The application-aware data reduction strategy and cost-accuracy tradeoffs may be more intelligently made based on knowledge of the application domain.

ML classification for BIM (with Solibri data)



Source: Minjung Ryu, „Machine Learning-based Classification System for Building Information Models “, Aalto CS Master thesis, 2020

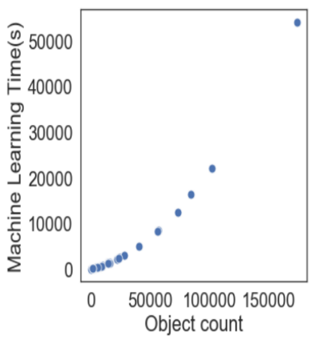
ML classification for BIM (with Solibri data)



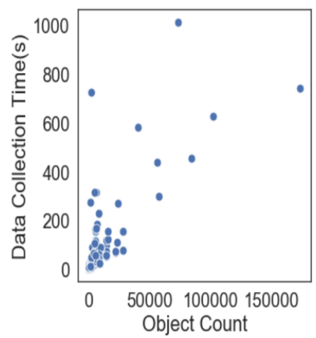
Source: Minjung Ryu, „Machine Learning-based Classification System for Building Information Models “, Aalto CS Master thesis, 2020

Initial results

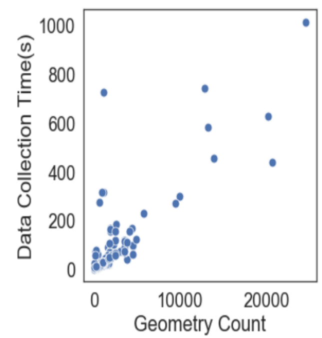
- Data set: 591 classification cases from 146 models
- Machines: AWS/Local with/out GPUs
- Different cases and settings



(a) Object count vs ML time



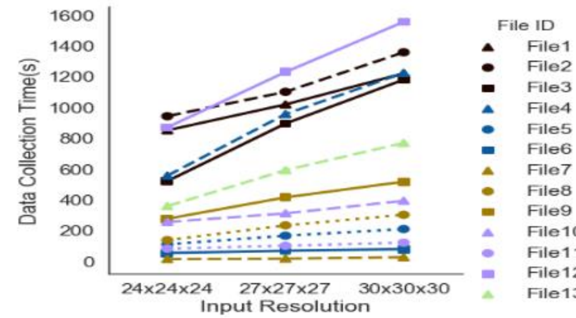
(b) Object count vs DC time



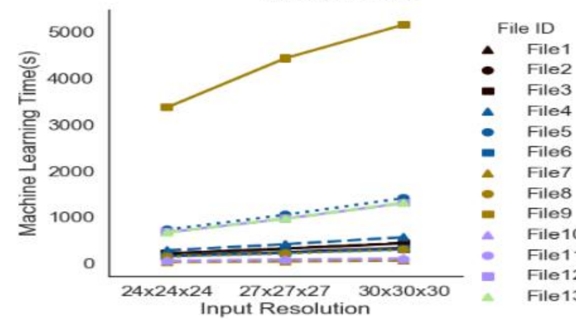
(c) Geometry count vs DC time

Figure 5: Impact of object counts on `DC time` and on `ML time`

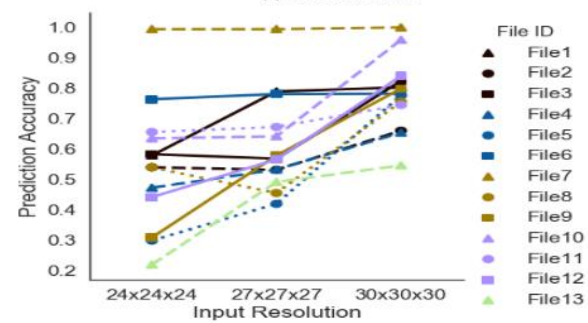
Reveal various relationships between types of data, extracting data resolution, machines and the accuracy of classifications



(a) Dim vs DC time



(b) Dim vs ML time



(c) Dim vs Accuracy

Aalto University
School of Science

Roadmap: Robustness, Reliability, Resilience and Elasticity (R3E) with QoA

QoA as an approach for ML optimization

- **A conceptual framework for defining Quality of Analytics (QoA)**
  - Metrics for services, data and ML models
  - Human-in-the-loop and domain expert integration
- **QoA as "contract"**
  - Leverage service contract and data contract models for QoA4ML
- **Monitoring and mechanisms for measuring QoA**
  - Measuring accuracy for data and models is challenging
  - Integration with data validation and model validation tools

QoA as an approach for ML optimization

- **Application-specific Resource Ensembles**
  - Resource ensembles are provisioned based on QoA
  - Containers and Kubernetes are key technologies for elastic edge-cloud platforms
- **QoA-aware and ML flow coordination**
  - Elasticity Primitive Operations
    - *Also leveraging quality data controllers and data governance processes in big data for ML?*
  - Elastic ML model serving platform-as-a-service
    - *Suitable for ensemble and federated ML?*

QoA as an approach for ML optimization

- **Models for predicting QoA**
  - Need to develop models capable of predicting QoA (of data and ML models)
- **Methods for adaptation and optimization**
  - Enable users to explore QoA tradeoffs such that they can inform application development and use
  - Integrate with different approaches to QoA-aware distributed computing

Conclusions

- **QoA can be used as a contract for Robustness, Reliability, Resilience and Elasticity (R3E) of ML**
  - Selected scenarios: training optimization, runtime ML model serving, out-of-distribution detection and optimization
- **Elasticity Engineering**
  - Can be a powerful techniques for achieving QoA in ML systems
- **Need real, complex ML systems for testing our ideas!**
- **Collaboration with FCAI is really important!**

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io